Class Size Analysis Appendices

Brian Jacob, University of Michigan

Rene Crespin, University of Michigan

CJ Libassi, University of Michigan

Susan M. Dynarski, University of Michigan

These appendices describe the sample used in the Michigan class size brief, Class Size in Michigan: Investigating the Risk of Being in Very Large Classes.

Data Sources
This project relies on several sources of administrative data on Michigan schools studies, all of which were originally collected by the Michigan Consortium of Education Research (MCER) and are currently maintained by the MCER and the Education Policy Initiative (EPI) at the University of Michigan.[1] Most importantly, the project utilizes student transcript data known as the TSDL (Teacher-Student Data Link), collected from individual school districts by the Michigan Department of Education (MDE) and Center for Education Performance Information (CEPI). This dataset includes course-level data on every class a student took with information on the school, teacher, subject, course name, semester and period.

The Sample
Our analysis focuses exclusively on data from the 2014-15 school year and students in grades 1, 7 and 9. We chose these grades to provide information on class sizes at different levels of the K-12 education pipeline. We begin will all students enrolled in these grades in 2014-15 in the master MDE-CEPI enrollment file who are attending schools with at least 10 students in that grade during this academic year. Because many students will attend more than one school during a year, we use the school attended for the greatest fraction of the school year as constructed by MCER. We take this sample of students and merge them to the TSDL file, only keeping observations from the TSDL for a single school per student. We exclude all observations from special needs schools as well as schools for which we cannot identify the urbanicity, poverty concentration, or charter designation of the school.[2]

Methods for identifying classrooms in the TSDL
In order to identify classrooms, we used the following five variables:

- school code
- nationally established subject code
- nationally established course code
- the educating entity's local course ID, which is used to uniquely identify local courses within a school
- the educating entity's course section ID, which is used to differentiate sections within the same local course ID (e.g., different periods during the day, different semesters, etc.)

In combination, these five variables specify a classroom. For example, School 123, English, English/Language Arts I, ENG9, 01SM1 represents School 123's, English Subject, English/Language Arts I Course, Titled English 9, Section 01 of semester 1.[3]

---

[1] For more information on MCER and EPI, see http://michiganconsortium.org/ and http://www.edpolicy.umich.edu/.
[2] These final restrictions result in dropping an extremely small number of cases.
[3] In the cases where a classroom had two teachers, each student would show up twice in that classroom with two identical records each matched with a different teacher ID. We count each unique student once in calculating class size even if they have multiple records for having multiple teachers.

In identifying unique classes, by necessity we dropped observations that were missing data on any of the five variables used to define the class (i.e., school, subject, course, local course ID, local section ID).[4] We also dropped all virtual courses because we wanted to capture class sizes of face-to-face courses only.

Analytic Choices and Sample Restrictions for Class-Level Data

Our goal is to create class size information for the "homeroom" class of each 1st grader and then for each of four core academic classes for every 7th and 9th grader (math, science, English and social studies).

For each student in 7th and 9th grade, we select one course per subject. If a student has more than one course in the subject, we choose what appears to be the most common course for that subject in their grade. For example, if a 9th grade student took both Algebra I and Consumer Math, we selected the Algebra I course. If a 7th grader took both "Science (Grade 7)" and Meterology, we would choose the course titled "Science (Grade 7)." If a student did not take any of the small set of courses we determined as "standard" for that subject in their grade, he or she was not included in the calculations for that subject (although he or she might still have valid class size information another subject).

Unlike 7th and 9th graders, many 1st grade students are in self-contained classrooms in which a single teacher instructs the students in multiple subjects. In the transcript files, these students have a record with the established subject code of NSS, which stands for "Nonsubject-Specific (prior-to-secondary)". According to the transcript documentation, this code stands for "courses that are not differentiated by subject area—that is, instances in which students enrolled in a grade-specified course and are taught various subjects throughout the day, rather than being enrolled in subject-specific courses."

Over 70% of 1st graders have at least one record with the subject code NSS. Some of these students also have records for what appears to be standard math or English classes. And virtually all of the students who did not have a record with the subject code NSS had at least one record with a standard math or English class code.[5]

If a student had more than one record with a common math or English record or the code NSS, we used the following decision rules to select a single classroom observation to use in the analysis. In the very small number of cases in which the choice of a single classroom would result in class size differences of more than one student, we simply excluded the student from the analysis. We mark students' English subject records as their homeroom if they have (i) an English, a math and a NSS class, (ii) an English and an NSS class, or (iii) an English and a math class. We mark students' math subject records as their homeroom if they have a math and an NSS class.

---

[4] We also dropped records with what appeared to be invalid local course IDs. In these cases, each student in the class was assigned a unique classroom ID, which does not allow us to connect students in the same classroom.

[5] The five most common English courses were Language Arts (grade 1), Reading (grade 1), Language Arts, English Language and Literature-Other, and Reading. The two most frequent math courses were Mathematics (grade 1) and Mathematics.

We then dropped entire classrooms in the following cases:

- Classrooms in which 100% of students were classified as limited English proficient or receiving special education services. We exclude these classes because they serve specialized populations and the sizes of these classes might not be comparable to "general" classes.

- Classrooms in which 100% of students were marked as having audited the course. We exclude these because we suspect that they were not "true" classes in the sense that one typically defines a class.

- Classrooms with fewer than 5 students or more than 55 students. We suspect that these were either data errors.

After all of the exclusions described above, we are still left with a number of students who have more than one record for a specific subject. The most common reason for this is that courses are listed multiple times depending on the term structure that the district uses. For example, a student who took a single Algebra 1 course with the same students in the same classroom, may have more than one record for that class if the transcript recorded the class as four separate one-quarter courses or two separate one-semester courses. Without a careful manual review of literally every transcript file in every school in the state, we cannot determine which records represent the same classroom in a different academic term.

Another possible reason for this is that students can take different courses that fall under the same subject and course code. Based on the variables we use to identify classrooms, for example, it may seem like a ninth grader is taking two Math Algebra 1 courses, but reading the local course ID and title actually shows us that the student is taking an Algebra 1 course and one Algebra 1 tutoring course. Again, without careful manual review of the variable combinations we cannot always tell which course is more representative of standard core classes for each student.

Only one observation per student is kept in the analysis, even though they may have class size data for multiple subjects or for multiple courses in the same subject (the two scenarios described above). We use the maximum class size experienced by each individual student over all of their final core subjects and observations. In practice, there is virtually no difference between the class size statistics obtained this way compared with using the median class size experienced by the student.

If a classroom is left with less than 50% of its students in the analysis sample, or less than 10 students in the classroom, after the aforementioned exclusion criteria, the classroom is dropped as there not sufficient data to report. In some cases, all classrooms in a school are dropped, resulting in the entire school being dropped. Similarly, if all schools in a district are dropped, an entire district might be dropped as well. Districts are excluded from the analysis if less than 50% of the students are represented in the analysis sample.

It is important to note that the class sizes are based on all students in a class, including students in grades other than our "target" grades (1, 7 and 9) and students who only attended the school for part of the school year. In contrast, the number of students we report for a specific grade and school is limited to students in grade 1, 7 or 9 for whom the school listed is the school the student attended for the majority of the school year.